Are Emergent Abilities of Large Language Models a Mirage?

Abstract

Recent work claims that large language models display emergent abilities, abilities not present in smaller-scale models that are present in larger-scale models. What makes emergent abilities intriguing is two-fold: their sharpness, transitioning seemingly instantaneously from not present to present, and their unpredictability, appearing at seemingly unforeseeable model scales. Here, we present an alternative explanation for emergent abilities: that for a particular task and model family, when analyzing fixed model outputs, emergent abilities appear due to the researcher's choice of metric rather than due to fundamental changes in model behavior with scale. Specifically, nonlinear or discontinuous metrics produce apparent emergent abilities, whereas linear or continuous metrics produce smooth, continuous predictable changes in model performance. We present our alternative explanation in a simple mathematical model, then test it in three complementary ways: we (1) make, test and confirm three predictions on the effect of metric choice using the InstructGPT/GPT-3 family on tasks with claimed emergent abilities; (2) make, test and confirm two predictions about metric choices in a meta-analysis of emergent abilities on BIG-Bench; and (3) show to choose metrics to produce never-before-seen seemingly emergent abilities in multiple vision tasks across diverse deep networks. Via all three analyses, we provide evidence that alleged emergent abilities evaporate with different metrics or with better statistics, and may not be a fundamental property of scaling AI models.